Homework 6

The purpose of this homework is to practice running parametric hypothesis tests and simple linear regression. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through gradescope by 11:59pm on Sunday October 13th.

As always, if you need help with any of the homework assignments, please attend the TA office hours which are listed on Canvas and/or ask questions on Piazza. Also, if you have completed the homework, please help others out by answering questions on Piazza.

Part 1: Parametric tests comparing two means - the t-test

Sleep or Caffeine for Memory revisited

On problem 3 of homework 4, you used a permutation tests for comparing two means. The description on the homework of that study was:

The consumption of caffeine to benefit alertness is a common activity practiced by 90% of adults in North America. Often caffeine is used in order to replace the need for sleep. One recent study compared students ability to recall memorized information after either the consumption of caffeine or a brief sleep (see Mednick et al., 2018

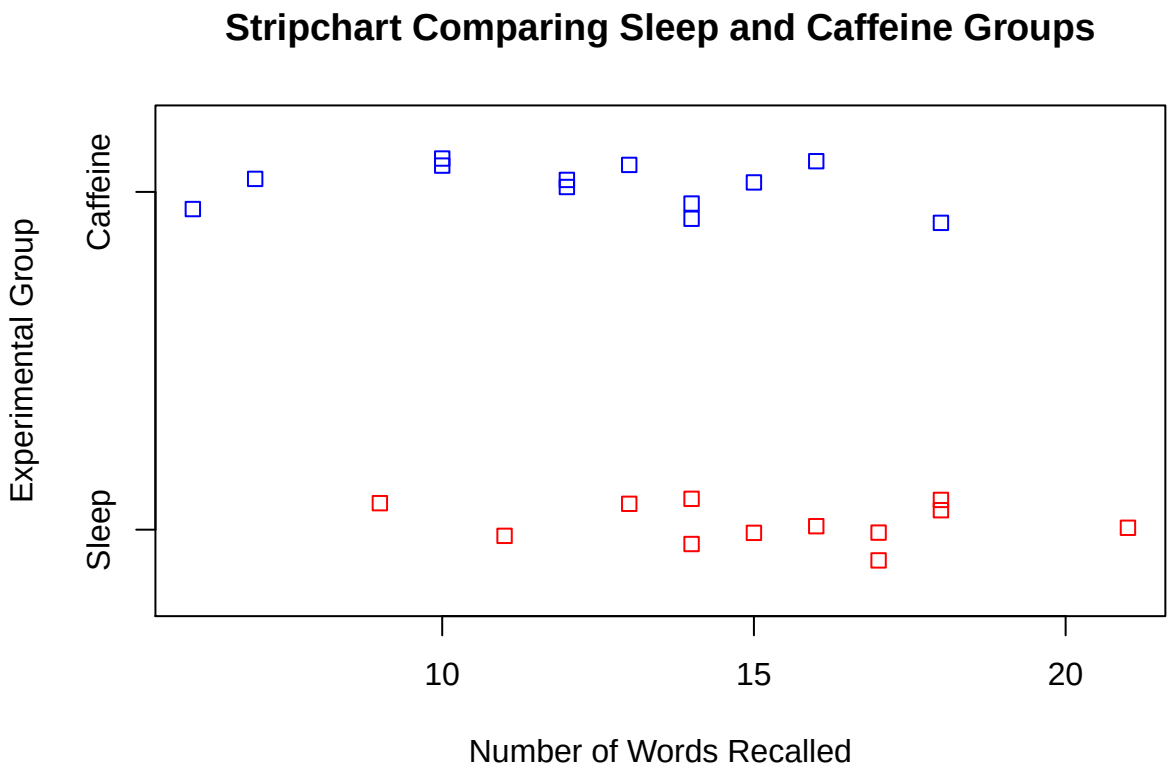
A random sample of 35 adults (between the ages of 18 and 39) were randomly divided into three groups and verbally given a list of 24 words to memorize. During a break, one of the groups took a nap for an hour and a half, another group was kept awake and then given a caffeine pill an hour prior to the testing, and a third group was given a placebo. The response variable of interest is the number of words participants are able to recall following the break.

Let's now run a parametric hypothesis test (t-test) to see if there is statistically significant difference in the mean number of words recalled between the group that got *sleep* and the group that got *caffeine*.

Part 1.0 (5 points) The data for the number of words recalled by members in each of the two groups is below. Start by creating a stripchart comparing the two groups. From this plot, does there appear there would be any major concerns running a t-test to analyze whether there are differences between the means of these groups?

```
sleep_condition <- c(14, 18, 11, 13, 18, 17, 21, 9, 16, 17, 14, 15)
caffeine_condition <- c(12, 12, 14, 13, 6, 18, 14, 16, 10, 7, 15, 10)

stripchart(list(sleep_condition, caffeine_condition),
           group.names = c("Sleep", "Caffeine"),
           ylab = "Experimental Group",
           xlab = "Number of Words Recalled",
           col = c("red", "blue"),
           method = "jitter",
           main = "Stripchart Comparing Sleep and Caffeine Groups")
```



Answer

Since the sample size for both populations is small (n = 12) for both, there are some concerns with using a t-test. The sample should either be large or normally distributed, but from the stripchart it appears more uniformly distributed than normally distributed.

In parts 1.1 to 1.5 you will now do the 5 steps to run a hypothesis test using parametric methods.

Part 1.1 (2 points) State the null and alternative hypotheses using words and symbols. Also describe the significance level is and denote it with the appropriate symbol.

Answer:

H_0 : There is no difference in the memorization ability of people who sleep versus people who consume caffeine.

$\mu_{sleep} - \mu_{caffeine} = 0.$

H_A : There is a difference in the memorization ability of people who sleep versus people who consume caffeine.

$\mu_{sleep} - \mu_{caffeine} \neq 0.$

Significance level: $\alpha = 0.05$. The significance level is the probability that we reject the null hypothesis when it is true (we conclude there is a difference in memorization ability when there is not). This is a Type I error.

Part 1.2 (8 points) Calculate a t-statistic for the observed sample and report the value. Based just at looking at the statistic value, does it seem that the results will end up being statistically significant? (hint: if this was a z value from a standard normal distribution, would it be statistically signficiant?).

```
(total_SE <- sqrt(var(sleep_condition)/length(sleep_condition)+
                    var(caffeine_condition)/length(caffeine_condition)))
```

```
## [1] 1.399405
```

```
(obs_tstat <- (mean(sleep_condition) - mean(caffeine_condition))/total_SE)
```

```
## [1] 2.143769
```

Answer

Based on the t-value it does seem that the results will be statistically significant. Since we are performing a two-tailed test, and the t-statistic is more than 2, then the density under the curve for both tails should be less than 0.05 (since two standard deviations on a standard normal distribution covers 95% of the curve). Thus p-value will be less than $\alpha = .05$ and the results will be significant.

Part 1.3 (10 points) Identify and plot the null distribution, and report the degrees of freedom. Also add red vertical lines to your plot at the observed statistic value(s) to indicate the amount of probability area in the tails of the null distribution that are more extreme than the observed statistic value.

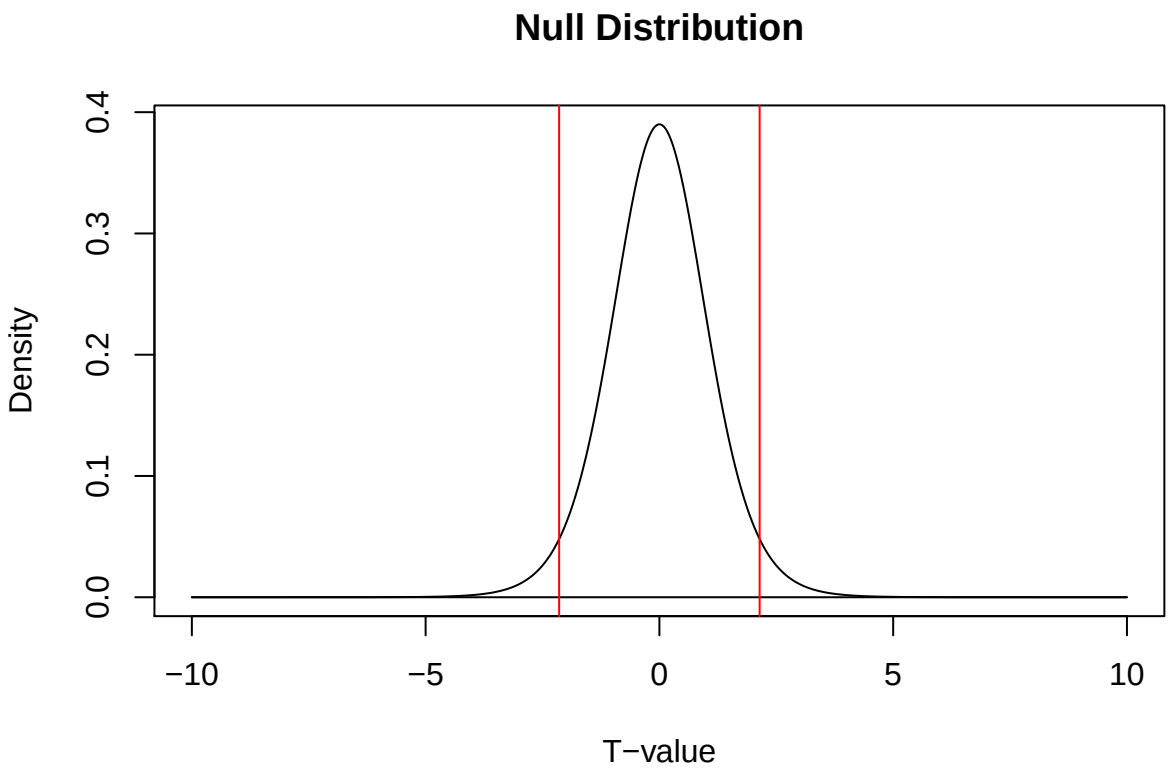
```
# sample sizes for the two conditions
n1 <- length(sleep_condition)
n2 <- length(caffeine_condition)

# estimate the degrees of freedom
(df <- min(n1 - 1, n2 - 1))
```

```
## [1] 11
```

```
# plot the null distribution with the observed statistic on it
x_vals <- seq(-10, 10, length.out = 1000)
y_vals <- dt(x_vals, df)

plot(x_vals, y_vals, type = "l",
     main = "Null Distribution",
     xlab = "T-value",
     ylab = "Density")
points(x_vals, rep(0, length(x_vals)), type = 'l')
abline(v = obs_tstat, col = "red")
abline(v = -obs_tstat, col = "red")
```



Answer

The degrees of freedom is 11.

Part 1.4 (7 points) Now calculate the p-value in the R chunk below.

```
(p_val <- pt(obs_tstat, df = df, lower.tail = FALSE) + pt(-obs_tstat, df = df))
```

```
## [1] 0.05524225
```

Part 1.5 (15 points) As discussed in class, Null Hypothesis Signifiance Testing is a hybrid of two different theories, namely Fisher's "significance testing" and Neyman and Pearson's "hypothesis testing" (to read more about this see this paper.

Please answer the following questions to interpret the results in light of these theories:

Based on **Neyman and Pearson's "hypothesis testing" paradigm**: 1) Are the results statistical significant at a significance level of $\alpha = 0.05$? 2) Does it seem likely you are making a Type I error here? 3) Does it seem likely you are making a Type II error here?

Based on **"Fisher's significance testing" paradigm**:
4) Does there seem to be a difference between these groups?

5) Which paradigm do you think best gets at what is truly happening?

Answers:

1) Since the p-value is 0.055 and greater than α, the results are not statistically significant according to the Neyman and Pearson paradigm, and we fail to reject the null hypothesis.
2) No, it does not seem likely, since we did not reject the null hypothesis.
3) It seems likely that we would be making a Type II error because we fail to reject the null, but the p-value is still low and close to the significance level.
4) Based on Fisher's significance testing paradigm, since the p-value is low (0.055), there is evidence against the null hypothesis and it seems that there is a difference between the two groups.
5) It seems that Fisher's paradigm gets best at what is happening. It gives a more accurate conclusion given existing scientific literature, and the data we have does give strong evidence against the null despite the p-value being more than α.

Part 1.6 (5 points) As we also discussed in class (and the previous homework) R has several built in functions to do parametric hypothesis tests. In particular, R has a built in function called `t.test(sample1, sample2)` that takes two samples of data and runs a t-test on them. Use this function to compare the sleep and caffeine groups and report the p-value (which should be slightly different from the one you got when you did your own t-test above). Also, describe whether your conclusions would be different using this built in function compared to when you ran your t-test above if you a) followed the Neyman-Pearson's hypothesis testing paradigm, and b) if you followed Fisher's significance testing paradigm. Finally, describe the reason why your p-value differs from the p-value returned by the t.test() function.

```
t.test(sleep_condition, caffeine_condition)
```

```
##
##  Welch Two Sample t-test
##
## data:  sleep_condition and caffeine_condition
## t = 2.1438, df = 21.894, p-value = 0.04342
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09699633 5.90300367
## sample estimates:
## mean of x mean of y
##     15.25     12.25
```

Answers:

The p-value using the t.test() function is 0.04342, which is less than the one calculated above. Now, according to the Neyman-Pearson paradigm, the results are statistically significant since the p-value is less than the alpha value, and we reject the null hypothesis (a different result from above). If we follow Fisher's paradigm there is also significant evidence against the null hypothesis as well since the p-value is small (the same conclusion as above).

The difference in the two tests is due to a change in the degrees of freedom. Since df is greater in this test, the null distribution is more normally distributed and there is less area under the null distribution curve at the tail ends, meaning the same t-value will result in a smaller p-value when the degrees of freedom is higher.

Part 1.7 (7 points): Based on your answer in part 2.6, modify the t-test code you wrote to produce results that are consistent with R's t.test function.

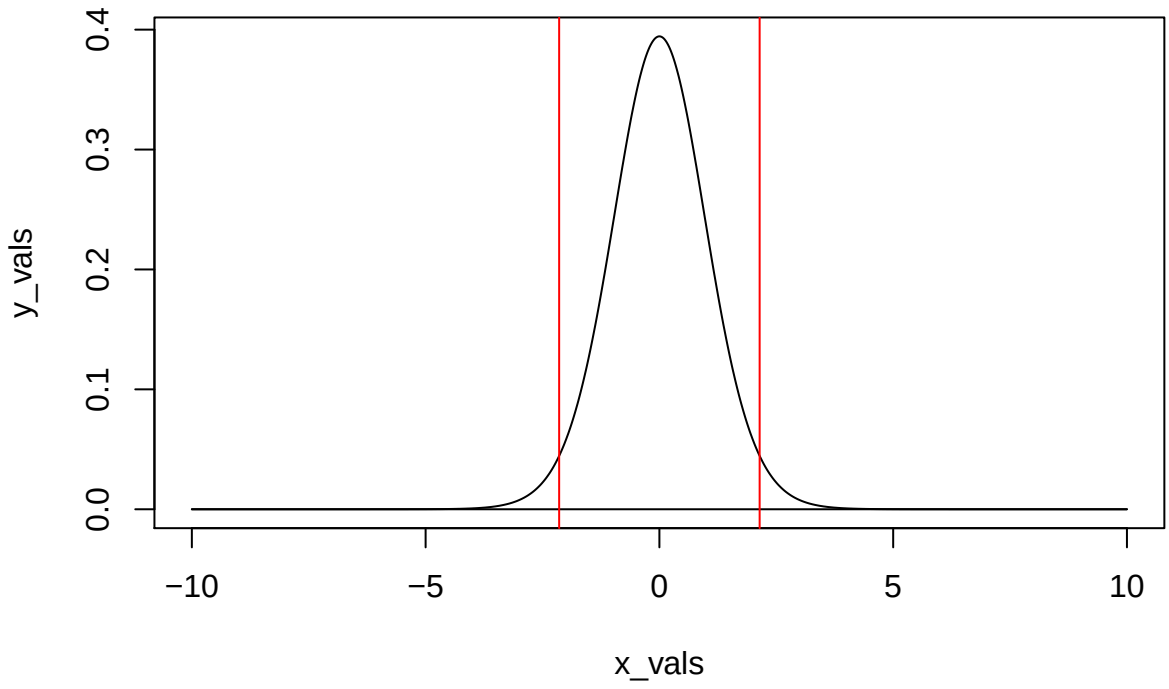
```
# sample sizes for the two conditions
n1 <- length(sleep_condition)
n2 <- length(caffeine_condition)

# estimate the degrees of freedom
(df <- 21.894)
```

```
## [1] 21.894
```

```
# plot the null distribution with the observed statistic on it
x_vals <- seq(-10, 10, length.out = 1000)
y_vals <- dt(x_vals, df)

plot(x_vals, y_vals, type = "l")
points(x_vals, rep(0, length(x_vals)), type = 'l')
abline(v = obs_tstat, col = "red")
abline(v = -obs_tstat, col = "red")
```



```r
(p_val <- pt(obs_tstat, df = df, lower.tail = FALSE) + pt(-obs_tstat, df = df))
```

```
## [1] 0.04341568
```

Part 2.1 (12 points)

In class we discussed that the mathematical derivation of different parametric tests are based on a set of assumptions/conditions (For actual derivations for population tests see this link)[https://pdfs.semanticscholar.org/6297/58de27161160c5ce051a6736c8b2004b42bc.pdf].

In practice, the inferences for a t-test are usually valid when a set of "rules of thumb" have been met. For a t-test, these rules of thumb are that either: the data looks relatively normal (i.e., doesn't have any large outliers), or alternatively, the sample size is n $>$ 30.

Let's do a simulation to see how robust the t-test is. Start by creating a null vector called `p_values` and then write a for loop that runs 10,000 simulations, where in each simulation you:

1) Draw a random sample of size n = 30 from an exponential distribution using the `get_sample(n)` function I have written. Save this sample in an object called `sample1`
2) Draw a second random sample also of size n = 30 using the `get_sample(n)` function. Save this sample in an object called `sample2`
3) Run a t-test and save the p-value for the t-test in the vector `p_values`

Once the for loop is done running, calculate the proportion of p-values that were less than $\alpha = 0.05$. Report whether it is the case that 5% or less of these p-values are indeed less than 0.05. Also plot a histogram of the p-values you collected and report the shape of this histogram.

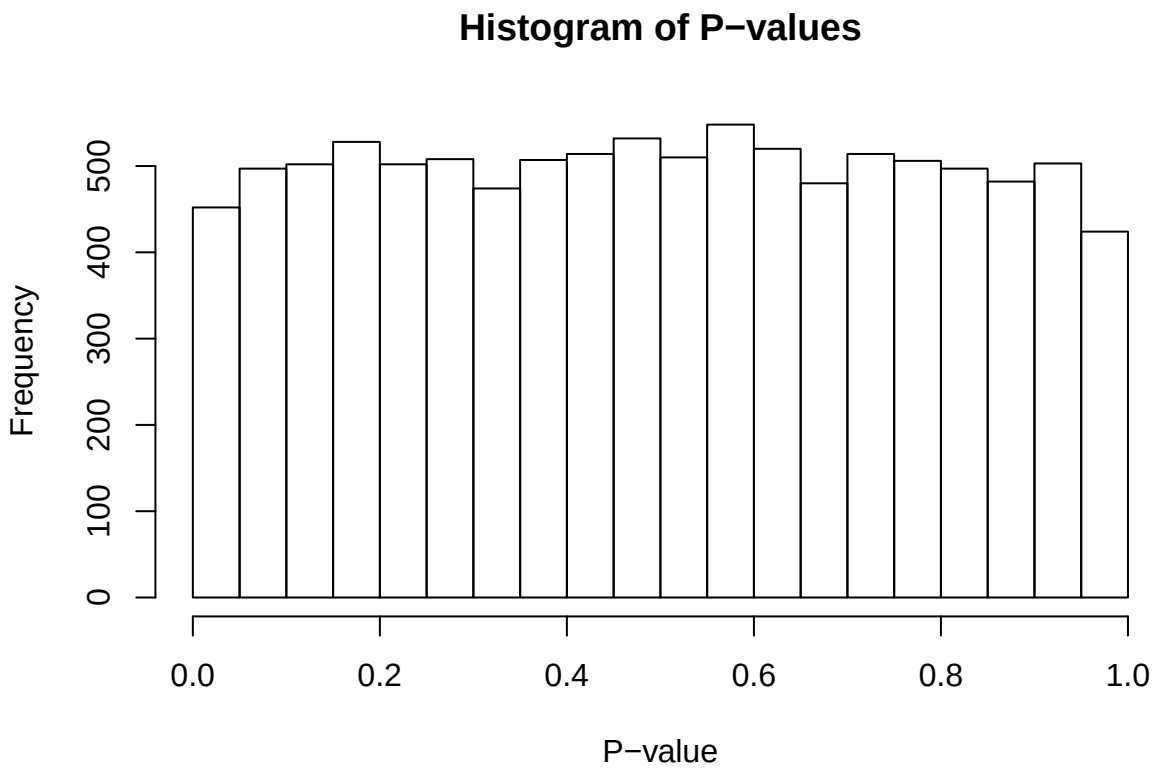
```r
set.seed(123)
# a function that returns a sample of data of size n (from an exponential distribution)
get_sample <- function(n){
  rexp(n)
}

# write a for loop that 10,000 time, gets two samples of size n
# and calculate the p-value
p_values <- NULL
for (i in 1:10000){
  sample1 <- get_sample(30)
  sample2 <- get_sample(30)
  p_values[i] <- t.test(sample1, sample2)$p.value
}


# see if the percentage of significant p-values is what is expected based on the alpha level
mean(p_values <= 0.05)
```

```
## [1] 0.0452
```

```r
#histogram
hist(p_values, main = "Histogram of P-values",
     xlab = "P-value",
     ylab = "Frequency")
```



Answers:

The proportion of p-values less than 0.05 is 0.0452, which is indeed less than 5 percent. The histogram of p-values has a uniform distribution.

Bonus part 2.2 (0 points)

Try changing properties of the simulation above to see if you can 'break' the t-test, i.e., if you can get more than 5% of p-values to be less than the signifiance level of $\alpha = 0.05$. In particular, try changing the sample size n, and the underlying distribution of the data (i.e., change the `get_sample()` function). For the changes you make, always keep n to be at least 10, and do not use any if statements in your `get_sample()` function.

Part 3: Simple linear regression

In 2000, the United States presidential election was between a Yale alumnus, George W. Bush who was the Republican candidate, and a Harvard alumnus Al Gore who was the Democratic candidate. There were also

a number of "third-party" candidates such as Princeton alumnus Ralph Nader who was the Green Party candidate, and Georgetown alumnus Pat Buchanan who was the Reform Party candidate.

The code chunk below contains data from the 2000 election for the state of Florida in a data frame called `florida data`. Each observational unit in this data frame contains information from the 67 counties in Florida including demographic information on each county as well as the votes received by each candidate in each county.

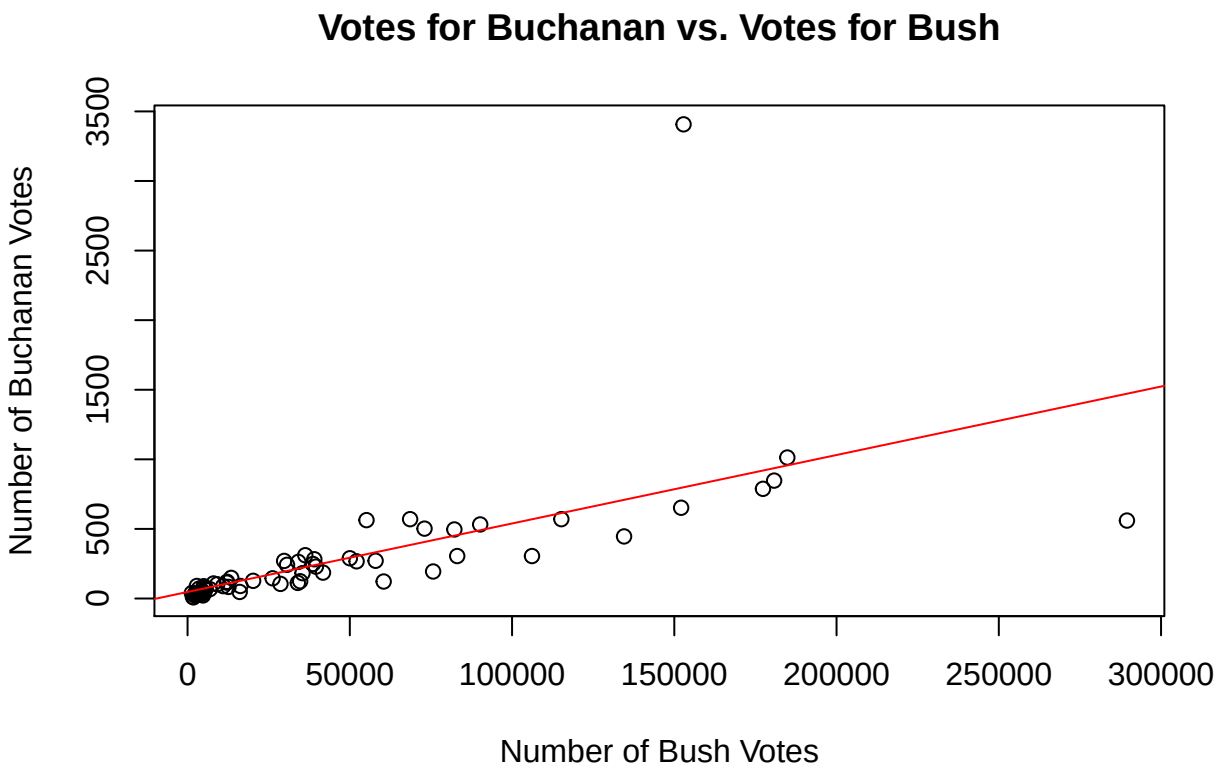
This the exercises below, you will use linear regression to look at the relationship between the votes that the Republican candidate *George W. Bush* received and the votes that the Reform candidate *Patrick Buchanan* received.

```
load('florida_vote_data_2000.Rda')
```

Part 3.1 (7 points):

Start the analysis by creating a scatter plot of the number of votes that Pat Buchanan recieved as a function of the votes that George Bush recieved. Then fit a linear model that can predict the number of votes Buchanan should receive given the number of votes that Bush received, and add the regression line to this plot in red.

```
plot(florida_data$Buchanan ~ florida_data$Bush,
     main = "Votes for Buchanan vs. Votes for Bush",
     xlab = "Number of Bush Votes",
     ylab = "Number of Buchanan Votes")
lm_fit <- lm(Buchanan ~ Bush, data = florida_data)
abline(lm_fit , col = "red")
```



Part 3.2 (7 points):

Now extract the coefficients from the linear model. In the space below report how many votes Buchanan is expected to get for every 1,000 votes Bush received, and how many votes the model predicts that Buchanan would have gotten if Bush had received 0 votes. Finally, write an equation that predicts the number of votes Buchanan should get as a function of the number of votes Bush received (make sure to use use $LaTeX$ for the proper notation).

```
coef(lm_fit)
```

```
##  (Intercept)         Bush
## 46.972816323  0.004920082
```

```
#Buchanan votes for every 1000 Bush votes
lm_fit$coefficients[2]*1000
```

```
##     Bush
## 4.920082
```

Answers:

To find the number of votes Buchanan is expected to get, we multiply the slope by 1000. So for every 1000 votes Bush gets, Buchanan is expected to get 4.92. If Bush had received 0 votes, Buchanan would have received $y = 46.97$ according to the model (the y-intercept).

Equation:

$\hat{y}$ = Predicted Number of votes Buchanan receives

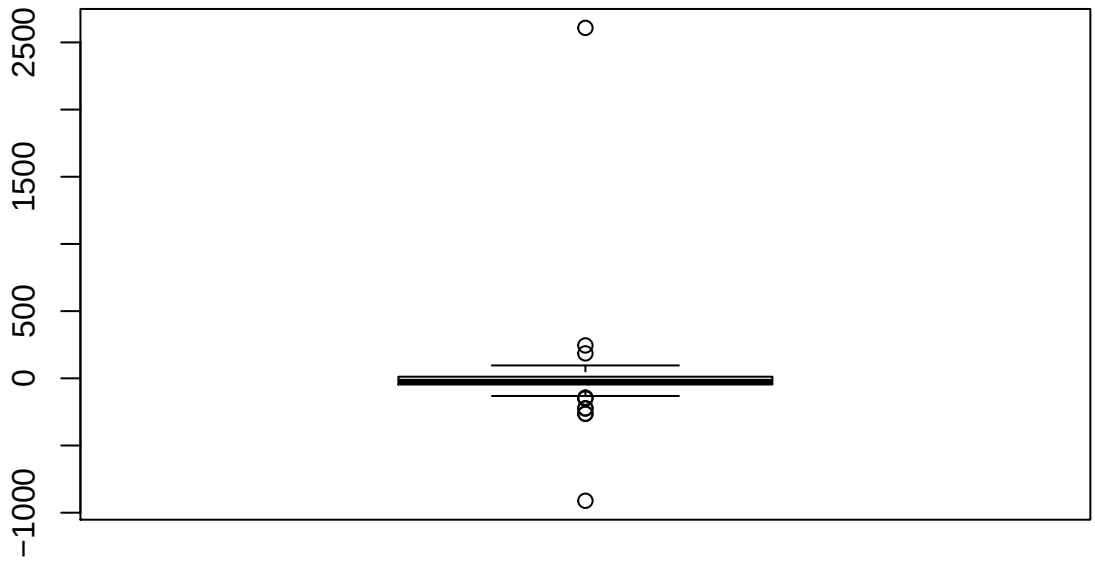
x = Number of votes Bush receives

$$\hat{y} = 0.00492x + 46.97$$

Part 3.3 (15 points):

From looking at the plot above, it should be clear that there is one extreme outlier. To see this more clearly, create a boxplot of the residuals of the model below. Then report: 1) what is the county that the outlier corresponds to, 2) how many votes Buchanan actually received in that county, 3) the predicted number of votes that Buchanan should have received for this county based on the regression model fit above for that county, and 4) the value of the residual for this county. Be sure to use the appropriate notation when reporting these numbers. Finally, use the Internet to come up with a reasonable explanation that could have led to this outlier (embedding images in the markdown document could be useful here).

```
# boxplot of the residuals
boxplot(lm_fit$residuals)
```




```
# county that is the outlier
#takes the county of the max residual
which.max(lm_fit$residuals)
```

```
## 50
## 50
```

```
florida_data[which.max(lm_fit$residuals),2]
```

```
## [1]   Palm Beach
## 67 Levels:   Alachua              ...   Washington
```

```
# actual number of Buchanan votes
florida_data$Buchanan[50]
```

```
## [1] 3407
```

```
# predicted number of Buchanan votes
lm_fit$fitted.values[50]
```

```
##       50
## 798.9876
```

```
# residual value
lm_fit$residuals[50]
```

```
##       50
## 2608.012
```

Answers

1) The outlier corresponds to Palm Beach county.
2) Buchanan received $y = 3407$ votes in Palm Beach.
3) The predicted number is $\hat{y} = 798.9876$.
4) The residual is $y - \hat{y} = 2608.012$.

According to the American Political Science Review, one possible explanation for this outlier is that the Palm Beach ballot, called a "butterfly ballot", was misleadingly designed such that Democrats mistakenly voted for Buchanan, thinking they were voting for Al Gore.

See Figure (on the next page) for a picture of the ballot.

Part 3.4 (7 points):

Suppose that Buchanan received exactly the number of votes predicted by the regression model, and the residual number of votes he received were intended to be votes for Al Gore. To examine the consequences of this, start by calculating the total number of votes Bush received and the total number of votes Gore received. Then add the residual number of Buchanan votes from the outlier county to the total number of votes that Gore received. Create an R Markdown table below and report these numbers. Would have this changed who got the majority number of votes (and hence who would have won Florida)?

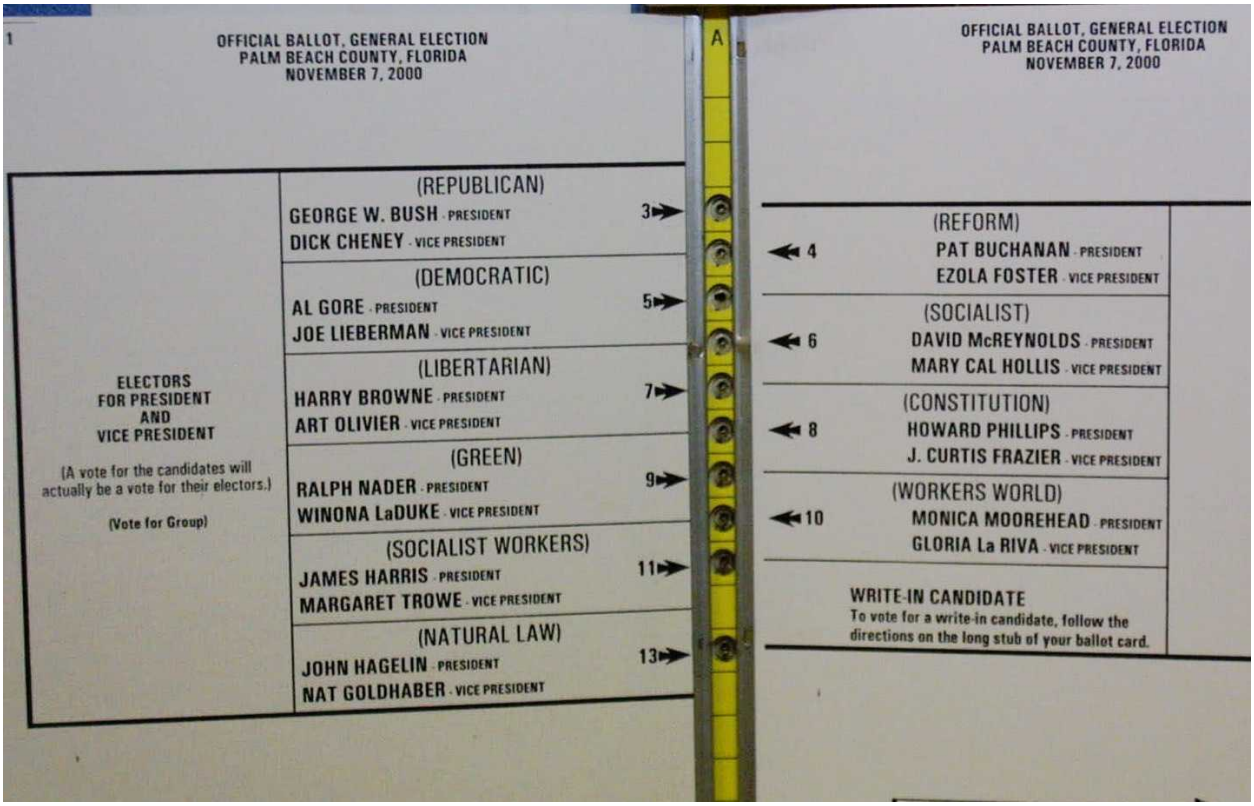


Figure 1: Picture of Butterfly Ballot

```
(bush <- sum(florida_data$Bush))
```

```
## [1] 2910078
```

```
(gore <- sum(florida_data$Gore))
```

```
## [1] 2909117
```

```
(gore_out <- sum(florida_data$Gore) + lm_fit$residuals[50])
```

```
##      50
## 2911725
```

Answers

Bush Total	Gore	Gore Total with Outlier
2910078	2909117	2911725.012387

Based on these numbers, if the residual votes were added to Al Gore's total, then Gore would have won Florida, not Bush.

Part 3.5 (2 points):

The United States uses the Electoral College system. In this system, the candidate who got the majority of the vote in a state wins all the Electoral College votes for that state (at least for most of the states in the US including Florida). Use the Internet to find the number of votes that Bush won the Electoral College